Project Report: Linear Regression - Insurance Dataset

Disclaimer

The following project report was written to satisfy the requirements of the final project for the Coursera Supervised Machine Learning: Regression course by IBM. It closely follows the materials outlined in the course including worked examples.

My code to produce figures as well as the figures are available in my github repository.

Objectives

For this exercise, I am training several linear regression models. The main purpose of this exercise is to see whether some features in a dataset (see description below) are able to **explain** the behavior of medical charges.

Description of Data Set

I am using the Medical Insurance data set available on Kaggle. The dataset consists of 1338 rows and 7 variable columns. The columns represent:

- Age, ordinal data with values between 18 and 64)
- Gender, categorical data consisting of 51% male and 49% female
- BMI (body mass index), numerical data between 16 and 53.1
- Number of children, ordinal data between 0 and 5
- Smoker, categorical data whether someone is a smoker
- Region, categorical data depending on the region ()
- Brief description of the data set you chose and a summary of its attributes.
- Insurance cost, numerical data of annual cost for insured person

The dataset does not contain any missing values.

Data Exploration

After having confirmed that the dataset does not have missing values, I decided to calculate basic statistics for numeric and categorical columns and also created a pair plot to assess whether there seem to be already clear relationships emerging.

I also decided to separate the data by whether the person is a smoker, because I am expecting that this makes a difference

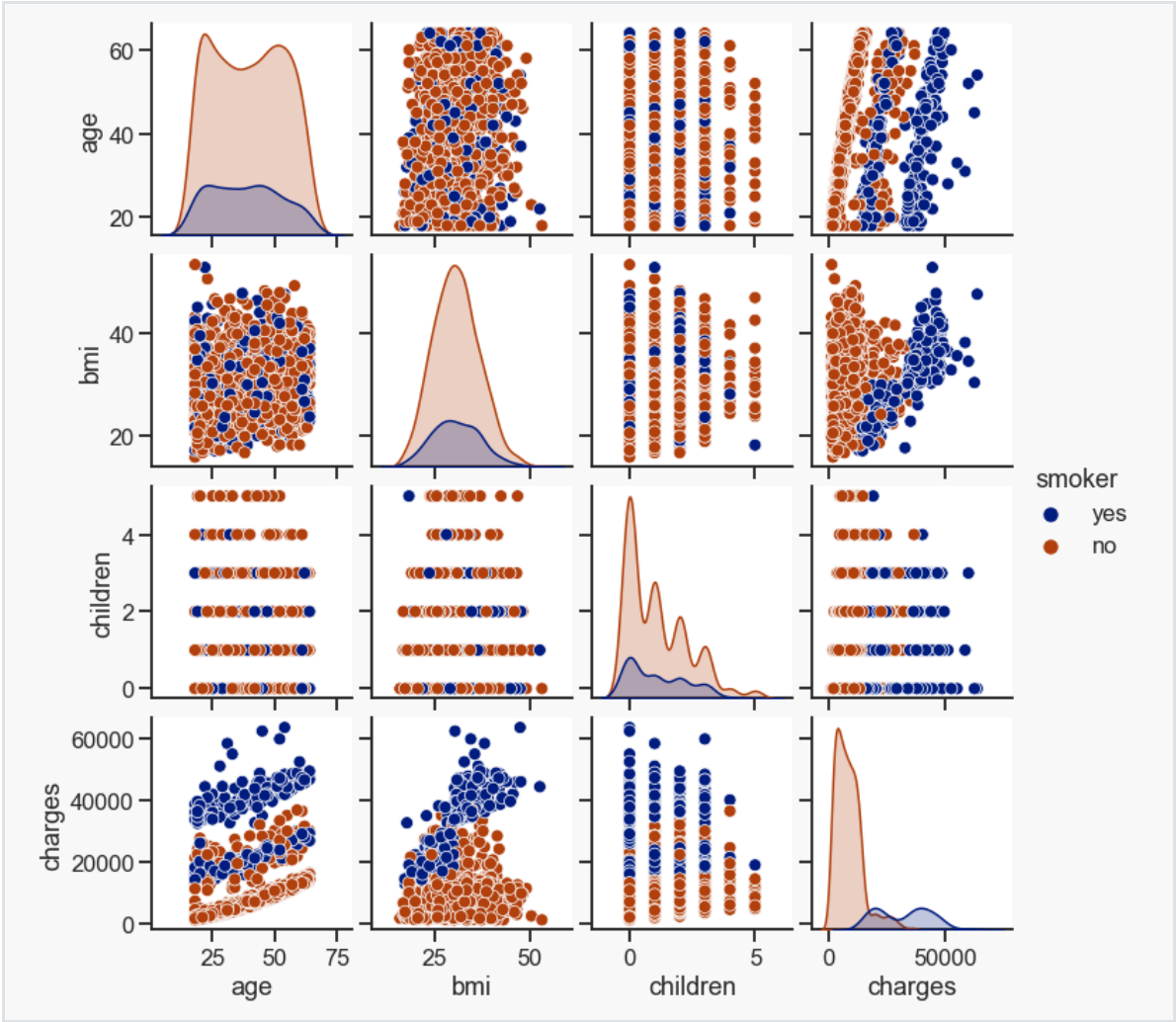

Figure 1: *Pairwise scatterplot of insurance dataset separated by smoker and non-smoker*

From Figure 1 can be seen that some columns, such as the number of children and the insurance charges are skewed and that there indeed seems to be a differing behavior in charges between smokers and non-smokers with smokers producing on average larger charges.

The skewness was calculated as

Column Name	Skewness
charges	1.51588
children	0.93838

Further investigation shows that the dataset is approximately balanced with respect to region and sex, but that there are many more non-smokers (n=1064) than smokers (n=274) in the dataset.

Preprocessing and Feature Engineering

The charges column was designated as the target and dropped from the feature list.

Based on the structure of the dataset, I decided to:

- Perform a log transform on the *charges* and *children* columns to reduce skewness
- Create polynomial features (for second degree polynomial) for numeric columns
- Perform one-hot encoding for ordinal columns

This was executed `transformer` objects for preprocessing in sklearn:

```
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.linear_model import LinearRegression

n_deg = 2 # degree of polynomial transformation for numeric features

numeric_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='median')),
    ('poly', PolynomialFeatures(degree=n_deg)),
    ('scaler', StandardScaler())])

categorical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='constant', fill_value='missing')),
    ('onehot', OneHotEncoder(handle_unknown='ignore'))])

preprocessor = ColumnTransformer(
    transformers=[
        ('num', numeric_transformer, numeric_features),
        ('cat', categorical_transformer, categorical_features)])

clf = Pipeline(steps=[('preprocessor', preprocessor),
                      ('regressor',  LinearRegression())])
```

Description of Models for Linear Regression

- Summary of training at least three linear regression models which should be variations that cover using a simple linear regression as a baseline, adding polynomial effects, and using a regularization regression. Preferably, all use the same training and test splits, or the same cross-validation method.

Model Selection

- A paragraph explaining which of your regressions you recommend as a final model that best fits your needs in terms of accuracy and explainability.

Key Findings

- Summary Key Findings and Insights, which walks your reader through the main drivers of your model and insights from your data derived from your linear regression model.

Outlook

- Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model adding specific data features to achieve a better explanation or a better prediction.

Grading Criteria

- Does the report include a section describing the data?
- Does the report include a paragraph detailing the main objective(s) of this analysis?
- Does the report include a section with variations of linear regression models and specifies which one is the model that best suits the main objective(s) of this analysis.

- Does the report include a clear and well presented section with key findings related to the main objective(s) of the analysis?
- Does the report highlight possible flaws in the model and a plan of action to revisit this analysis with additional data or different predictive modeling techniques?